8-2020

Analysis of the Functional Relationship of Protein Kinase Families Using Phospho-Proteomics Data

David A. Parra Peña
The University of Texas Rio Grande Valley

ANALYSIS OF THE FUNCTIONAL RELATIONSHIP OF PROTEIN KINASE FAMILIES USING PHOSPHO-PROTEOMICS DATA

A Thesis

by

DAVID A. PARRA PEÑA

Submitted to the Graduate College of
The University of Texas Rio Grande Valley
In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

August 2020

Major Subject: Computer Science

ANALYSIS OF THE FUNCTIONAL RELATIONSHIP OF PROTEIN KINASE FAMILIES USING PHOSPHO-PROTEOMICS DATA

A Thesis
by
DAVID A. PARRA PEÑA

COMMITTEE MEMBERS

Dr. Marzieh Ayati
Chair of Committee

Dr. Zhixiang Chen
Committee Member

Dr. Andres Figueroa
Committee Member

August 2020

ABSTRACT

Parra Peña, David A., Analysis of The Functional Relationship of Protein Kinase Families Using Phospho-Proteomics Data. Master of Science (MS), August 2020, 36 pp., 1 Table, 7 figures, 25 references, 21 titles.

As cancer research advances, Mass-spectrometry based proteomics is becoming a widely used technique for proteome characterization. Phosphoproteomics is a specific type of proteomics that characterizes proteins with the reversible post-translational modification of phosphorylation (PTM), which has allowed the identifications of thousands of phosphorylation sites. These phosphorylation sites, also known as substrates, are known to interact with a protein type named kinases. Studies have shown that abnormal phosphorylation activity is related to cancer diseases. Moreover, these kinases are divided into families, based on the similarity of their catalytic domain, as this part of their amino acid sequence determines a large part of what their functions are. In this work, propose 2 new methods to assess the relationship of kinases based on the correlation of the phosphorylation pattern of their substrates. Using these metrics, we cluster the kinases and analyze their inter-family interactions.

DEDICATION

I would like to dedicate this work to my family, who have been with me in every single step of this journey. Without their unconditional love and support none of this would be possible. To my friends who also shared this process with me, thank you.

ACKNOWLEDGMENTS

I will always be grateful to Dr. Marzieh Ayati, chair of my dissertation committee, for all her mentoring and advice. She always encouraged to pursue research in the area of Bioinformatics through her infinite patience and guidance. From data processing, solution design to the editing of this manuscript, she always extended a helping hand. My thanks go to my thesis committee members: Dr. Z, and Dr. Figueroa, for their flexibility and support, making my thesis defense possible. Also, I would like to acknowledge Daniel Acevedo, a student colleague, for his collaboration in the design of the methodology of this work, thanks.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Overview

Mass-spectrometry based proteomics is becoming a widely used technique for proteome characterization. Proteomic analysis of 3D cellular models and single-cell systems provides a method for correlating cellular heterogeneity and patient-specific responses to chemotherapy drugs, resulting in better cancer treatments with less side effects [1]. Phosphoproteoimcs is a specific type of proteomics that characterizes proteins with the reversible post-translational modification of phosphorylation (PTM), which has allowed the identifications of thousands of phosphorylation sites [2]. These phosphorylation sites, also known as substrates, are known to interact with a protein type named kinases. Studies have shown that abnormal phosphorylation activity is related to cancer diseases [7]. Moreover, these kinases are divided into families, based on the similarity of their catalytic domain, as this part of their amino acid sequence determines a large part of what their functions are.

In this work, we analyzed 3 different phosphoproteomics datasets with the goal of obtaining a better understanding of how these kinases interact in term of their respective families. In chapter 3, we describe the methods that we used to cluster the kinases, and in chapter 4, we present the result of this analysis.

CHAPTER II

REVIEW OF LITERATURE

Kinases and Phosphorylation

Protein Kinases are a class of enzymes in charge of regulating many different biological events. Protein phosphorylation is one of the regulatory mechanisms in which a kinase attaches a phosphate group (PO_4) to its target protein, allowing it to change conformation when interacting with other molecules (Figure 1). This mechanism is involved in many cellular processes such as protein synthesis, cell division, signal transduction, cell growth and development, and aging as many enzymes and receptors are activated and deactivated via phosphorylation/dephosphorylation events due to specific kinases and phosphatases. Moreover, kinases are known to regulate the majority of cellular pathways, particularly those involved in signal transduction. Being one of the most common Post-translational modifications (PTMs), protein phosphorylation is involved in the regulation of multiple biological processes and overexpression of kinases. When regulatory mechanisms are mutated or defective, the signaling pathways of kinases become dysregulated and/or abnormally active, being this the basis for oncogenesis for multiple tumors [8].

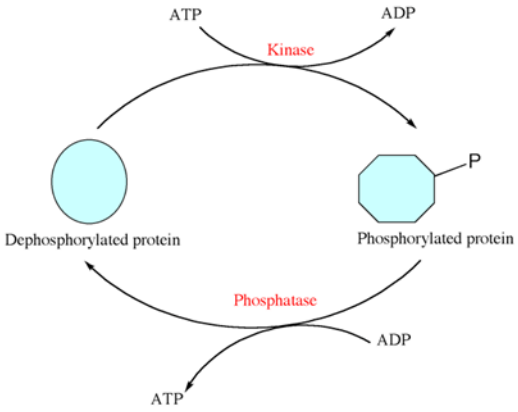


Figure 1. Phosphorylation process

Kinase phosphorylation on Cancer

Kinase over expression has been found in cancer related diseases. Therefore, understanding their interactions is key in the development of drugs in order to maximize the therapy's effect as well as minimize collateral damage in patients [6]. Studies have shown that abnormal phosphorylation activity has been found in Cancer diseases [8]. In breast cancer, therapies have developed to inhibit CDK4/6 which then expanded to other members of the CDK family, as they are known to regulate cell progressions in different phases of the cell cycle. In particular, CDK2 is a target of several candidate cancer. The inhibition of CDK2 however, has shown not to prevent cancer cell growth due to CDK redundancy [10]. Other kinases targeted in the treatment of breast cancer arePI3K/AKT and mTOR (PAM) because the inhibition of their pathway has shown to be beneficial as the PAM pathway has been estimated to be in as frequent as 70% of breast cancers [11]. Ovarian cancer is reported to have the Mirk/dyrk1B gene upregulated as it mediates cancer cell survival by decreasing the toxic ROS levels [12].

Human Kinome

Kinases are classified based on their sequence similarities into families, each containing subsets denoted as subfamilies. The Hanks and Hunter classification scheme is rooted in the catalytic domain (also known as kinase domain) phylogeny which enabled them to reveal conserved features of the domain [3]. Further additions to the scheme where made by the KinBase project, which utilized knowledge of sequence similarity and domain structure foreign to the catalytic domain, as well as known biological functions and gross similarity in these functions of kinases across organisms [4]. While these schemes describe the kinases as "members of a team with matching uniforms", it does not encompass their performance in diseases. In this project, we are trying to use the mass spectrometry-based phosphoproteomics data to cluster the kinases. For this purpose, we use the phosphorylation of the substrates of kinases to cluster the kinases based on their activities. Our hypothesis is the phosphorylation of kinase's substrates reflect the activity of the kinases and can capture the collaborative kinases better that the traditional sequence-based clustering.

CHAPTER III

METHODOLOGY

Data description

Phosphoproteomic data

The data utilized is comprised of 3 datasets, 2 for breast cancer and 1 for ovarian cancer derived from proteomic and phosphoproteomic profiles. The first set consists of mass-spectrometry based proteomics for breast cancer patient-derived xenografts consisting of 56874 phosphosites in 24 breast cancer PDX models generated by the isobaric tag method for relative and absolute quantification (iTRAQ Fold Change) [13]. For the purpose of this study, the observations missing intensity values in fifty percent or more of their samples were removed, resulting in intensity data for 34980 phosphosites. Furthermore, the second set of breast cancer data is composed by mass-spectrometry based phosphoproteomics analysis of TCGA breast cancer samples conducted by The NCI Clinical Proteomic Tumor Analysis Consortium [15]. This collection consists of 62679 phosphosites in 111 tumors, which resulted in 24704 after filtering. The third set of data of this study data is composed by mass-spectrometry based phosphoproteomics of ovarian HGSC tumors characterized by the Cancer Genome Atlas, containing 24429 phosphosites in 69 tumors [16]. Following the same filtering procedure in the previous descriptions, 6490 phosphosites remained in data.

Kinase-Substrate association data

In order to map the phosphosites (also known as substrates) in the experimental data to their respective kinase partners, a compilation of known kinase-substrate interactions was extracted from the PhosphoSitePlus database [17]. In conjunction to the kinase-substrate association data, the KinBase database for the human kinome was used to verify the correct name of the kinases and unify the data, as proteins usually have alternate names, as well as generate the kinase family groups or clusters. After matching the phoshophosites to their respective kinase for the first breast cancer dataset, kinases with less than 2 substrates where removed from the data, resulting in 1952 observations, with 117 kinases. Moreover, with the same filtering process over the second breast cancer data, 1458 observations with 103 kinases. Finally, 905 observations with 70 kinases were found on the ovarian cancer data set after processing.

Perturbation data

In order to validate our clustering method, we use an independent dataset. Phosphopeptide quantification data and kinase-perturbation benchmarking data containing data files for 80 labeled conditions [14].

Assesment of the relationship between protein kinases

As discussed before, the objective of this study is clustering kinases by correlating the phosphorylation pattern of their substrates and compare them with family-cluster perspective. In the following sections, we describe two methods that we used to quantify the relationships between kinases.

Group Correlation Score (GCS)

This method is introduced to evaluate the relationship of a given pair of kinases based on the correlation of phosphorylation of their substrates while also considering those substrates that are phosphorylated by both kinases (i.e. shared substrates) as well as the number of substrates that both kinases correlate.

Let us denote $k_n = \{k \mid k \in K\}$ as a kinase in the set of kinases K.

Let $S_{k_n} = \{S_1, S_2, S_3 \ldots S_m\}$ be the set of substrates with which k interacts.

For a given pair of kinases k_i and k_j, the Group Correlation Score (GCS) is computed as:

$$GCS(k_i, k_j) = \frac{\sum \left|\rho_{S_{k_i}, S_{k_j}}\right| [\rho \geq \alpha]}{n_\rho * \sigma_\rho} + \frac{\left|S_{k_i} \cap S_{k_j}\right|}{\left|S_{k_i}\right| + |S_{k_j}|}$$

Where,

$\left|\rho_{S_{k_i}, S_{k_j}}\right|$ is the absolute value of Spearman's rank correlation coefficient for phosphorylation of two given substrates.

$\left|S_{k_i} \cap S_{k_j}\right|$ is the cardinality of the intersection of between the set of substrates.

$\left|S_{k_i}\right|$ and $|S_{k_j}|$ are the cardinality of the set of substrates.

α is the cut-off threshold value for the correlation ρ.

n_ρ is the number of correlations greater than or equal to α.

And σ_ρ is the standard deviation of the correlations that passed the cut-off point.

Correlation of the kinase features (KFC)

In this method, we reduce the dimensions of phosphorylation of substrates of a kinase using singular value decomposition (SVD). For this purpose, for a given kinases, we pick the top two largest eigen vectors of the phosphorylation data of substrates of the kinase.

Let $M_k$be the data matrix of the substrates for a kinase with dimensions $m\ x\ n$. M_k is factorized as:

$$M_n = U\Sigma V^T$$

Where, Σ is an $m\ x\ n$ rectangular diagonal matrix. U and V^T are both unitary matrices, of dimensions $m\ x\ m$ and $n\ x\ n$ respectively.

V is found by transposing V^T from which the first two vectors are extracted as v to compute the feature F_k as:

$$F_k = M_k * v$$

For a give pair of kinases k_i and k_j, the Spearman's correlation is computed using their features F_{k_i} and F_{k_j} as KFC $\left(F_{k_i}, F_{k_j}\right) = \rho_{S_{k_i}, S_{k_j}}$.

RV Coefficient (RVc)

The RV coefficient is a multivariate generalization of the squared Pearson Correlation Coefficient which measures the relationship between two sets of variables. The principle of this method is that two sets of variables are perfectly correlated if there exists an orthogonal transformation such that the sets overlap. In this study, the adjusted RV coefficient is used [18].

Correlation metrics to distance conversion

Let d be the distance between a pair of kinases k_i and k_j, and $CM(k_i, k_j)$ as any of the previous correlation metrics explained. These are converted to distance with:

$$d = \frac{1}{CM(k_i, k_j) + 1}$$

For the Group Correlation Score $GCS(k_i, k_j)$ and the $RV\ coefficient$, and

$$d = \sqrt{1 - CM(k_i, k_j)}$$

For the correlation of the kinase features KFC.

Clustering Kinases

Traditionally, protein kinases are separated into families and subfamilies based on the amino acid sequence of their catalytic domains. One of the main objectives of this work, is to use the phosphoproteomics data to generate clusters of kinases and assess the statistical significance of family occurrences within a cluster.

Hierarchical clustering

While there exist many different clustering algorithms with different advantages and disadvantages, for the sake of simplicity, agglomerative hierarchical clustering was chosen as it is widely used in the area of computational biology. In addition, this approach produces a dendrogram which depicts the similarities and branching between groups in a data cluster, making the process of cluster number selection straightforward (Figure 2). While we tested the

partitioning of clusters in the range from 5 to 12, only the results for k=7 was reported as it seemed as a good medium.

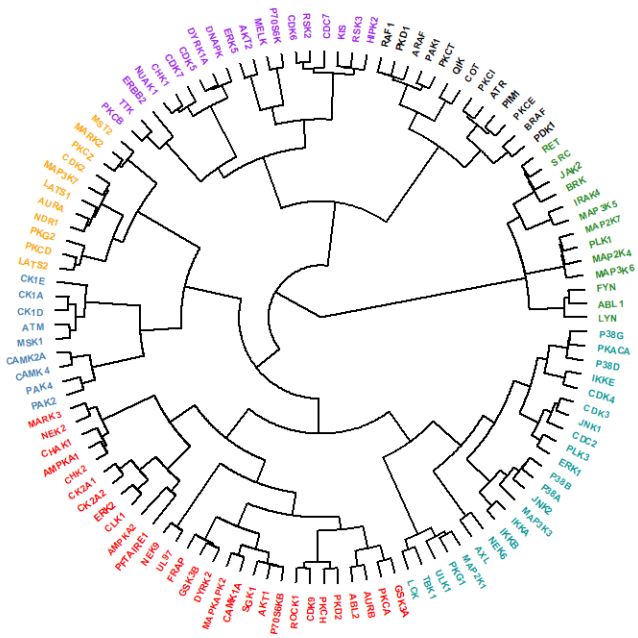


Figure 2. Phylogenetic tree of the phospho-clusters for Breast Cancer 1 from the GCS metric

Analysis of Kinase Clusters

Sequence-Based Clusters (Kinase Families) vs Phospho-Based Clusters

Protein kinases are grouped into families, based on the similarity of their sequence and catalytic domain. The hypothesis is that, while kinases with similar catalytic domains may participate together in the phosphorylation process, these may be interacting with other kinases that do may belong to different families. To test this hypothesis, the cluster overlap significance between the family groups and the clusters generated from the experimental phosphorylation data is computed. To evaluate the significance of the existing overlap between the clusters and

the family groups, we use the hypergeometric cumulative distribution function. In addition, in order to quantify the similarity of these phospho-clusters, we use Normalized Mutual Information (NMI) given by the formula:

$$NMI(Y, C) = \frac{2 * I(Y; C)}{H(Y) + H(C)}$$

Where, Y is the family class labels, C is the cluster labels, $H(.)$ is the entropy, and $I(Y; C)$ is the mutual information between Y and C.

Family Interaction Assessment

We also assess the collaboration of kinases from different families. For this purpose, we evaluate the number of kinases from different families that are grouped together in phospho-based clusters. We calculate a multivariate hypergeometric probability using the threshold $p \leq .025$.

Kinase-Substrate Enrichment Analysis (KSEA)

To further explore the functional relationship of kinases withing the phospho-based clusters, we used an independent dataset to measure the activities of kinases that are clustered together. For this purpose, we perform KSEA across the 80 conditions represented by the perturbation dataset [14]. KSEA systematically infers the activation of the pathways for a given kinase, providing a method for the systematic profiling of kinase pathway activities [19, 20]. The kinase's normalized score is calculated as follows:

$$score = \frac{(\bar{s} - \bar{p})\sqrt{m}}{\delta}$$

Here, $\bar{s}$ denotes the mean log2(FC) of know phosphosite substrates of the given kinase, $\bar{p}$ represents the mean log2(FC) of all phosphosites in the dataset, m denotes the total number of phosphosite substrates identified from the experiment that annotate to the specified kinase, and δ denotes the standard deviation of the log2(FC) across all phosphosites in the dataset [19].

CHAPTER IV

RESULTS

Overlap significance between the family groups and the clusters

The catalytic domain of kinases is relevant to the phosphorylation mechanism as it is the region that interacts with its substrates to cause an enzymatic reaction. However, this does not provide a broad enough picture that accurately represents the phosphorylation activity in the experimental data. For each data file, we intersect the sets clusters provided by each kinase relationship metric and the family groups to evaluate if a given family is strongly represented within a cluster. In addition, we compute the interaction of the families within a cluster to provide further analysis (The overlap significance values use $p < .05$ as a threshold).

Starting the analysis of the first breast cancer cluster sets (BC1), the Group Correlation Score (GCS) clusters report 11 significant intersections with 8 distinct families (NMI = 0.2598), having 4 clusters overlapping with 2 families each (see Figure 3). The set of clusters generated with the KFC method produced 11 significant overlaps with the 9 of the family groups (NMI = 0.1553). As well as GCS, this method shows 3 clusters that overlap significantly with 5 families. Continuing, the clusters from the RV coefficient had the same number of significant overlaps with the family groups as the previous method, with a total count of 11 (NMI = 0.2019). From the graph, we can observe that it presents 4 clusters that interconnect with at least 2 families each. Our data reports some combinations of overlaps appearing in at least two of the clustering

methods. For instance, the GCS set contains a cluster that overlaps with both the STE and TK families; such combination is also found in the RV set of clusters. Similarly, the KFC set, and the RV set present a cluster each that overlaps significantly with the AGC and CMGC families.

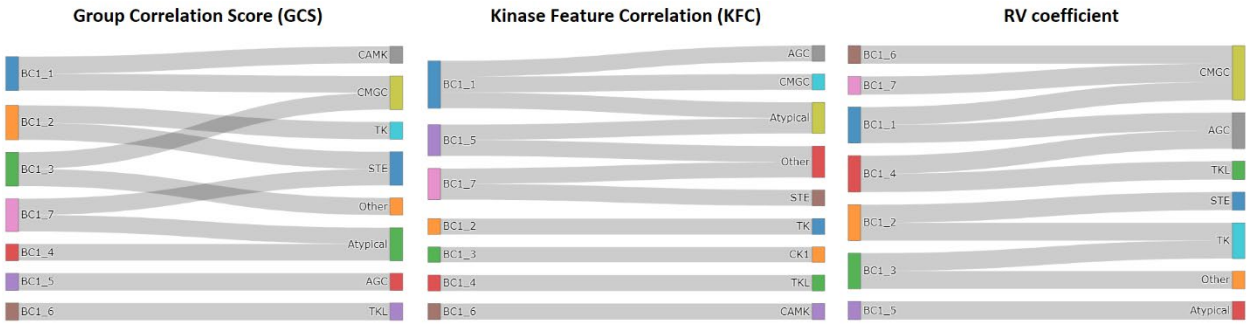


Figure 3. Cluster Overlap Significance between the clusters from BC1 and the Family groups

Following with the second breast cancer cluster sets BC2, GCS clusters report 7 significant overlaps connecting 6 of the clusters to 5 different families (NMI = 0.1534), with the combination of TKL and AGC at cluster 5. Next, the KFC clusters are found to have 8 significant overlaps, presenting the combination of families AGC with CMGC again, as well as well as Other and TK both overlapping with 2 clusters together (NMI = 0.2018). The RV coefficient method reports 9 significant overlaps between 6 of the 7 clusters from the set and 7 families (NMI = 0.2735). In this dataset we found overlaps of AGC and TKL in a cluster of the GCS set, as well as in the RV coefficient set. It is also worth noticing that the KFC method also reported AGC and CMGC overlapping significantly with a cluster (Supplementary Figure 1).

Continuing with ovarian cancer cluster set (OC), the clusters from the GCS method contain 11 overlaps with a significant number of interconnections between 4 of the clusters with

5 families (NMI = 0.4117). Likewise, KFC clusters have 10 significant overlaps while providing 8 links interconnecting 4 families (STE, TKL, TK, Other) to 6 clusters (NMI = 0.1535). Finally, RV coefficient produced 9 significant overlaps, while providing 7 links connecting 6 families to 4 clusters, with a NMI value of 0.2969 (Supplementary Figure 2).

To summarize the previous results, the analysis shows that in most clusters, at least one family will be significantly present while sharing membership with other families. In addition, results show that the kinase from different families might functionally be relevant. Moreover, we show that might a given cluster might even have more than one family strongly represented as each of the methods reported at least 1 cluster overlapping with at least 2 families. However, this does not mean that 2 clusters with a given combination of families are going to contain the exact same kinases.

Overlap significance between the different sets of clusters

In the past section we showed that in the different cluster sets, there are some kinase families that are present together in across the clusters, suggesting the existence of an underlying structure present in the data. We used an independent dataset (BC2, and OC) to assess the reproducibility of this collaborative patterns among kinase families. For this purpose, we identified the kinase clusters using BC2 and OC datasets, and we measure the intersection of these clusters with cluster identified based on BC1. For consistency, sets of clusters from different datasets are only analyzed together if they were generated using the same kinase relationship metric. These results are presented in Figure 4 and Figure 5.

As seen in the figure 3, there are 7 significant overlap across 6 clusters from BC1 and 5 from BC2S. In contrast, KFC clusters report 11 significant overlaps across all the clusters from both datasets. Lastly, the clusters from the RV coefficient provide 8 significant overlaps that link the 7 clusters of BC1 with 5 Clusters of BC2. These results suggest that the kinases of BC1 and BC2 have some significant groups in common across them, and the clusters identified in BC1 are reproducible. In addition, KFC clusters reported AGC and CMGC connection to a single cluster in both datasets, while RV coefficient reported the AGC-TKL combination on both datasets. BC1 and OC sets of clusters computed with the GCS metric report 6 overlaps across 4 clusters each. The KFC cluster sets report 7 links between 4 clusters from BC1 with 6 clusters from OC. RV coefficient cluster sets show 6 links across 4 clusters from BC1 and 4 from OC.

Between the BC2 and OC cluster sets for GCS metric, we report 5 significant cluster overlaps with 4 of them from BC2 interconnecting to 3 from OC. Consequently, KFC set reported 6 significant overlaps between BC2 and OC across 4 clusters each. Continuing with the RV coefficient clusters, we find 5 significant overlaps with linkage between 4 clusters per set. No common pair of families sharing a cluster across the data is reported. The result of this analysis is presented in Supplementary Figure 3.

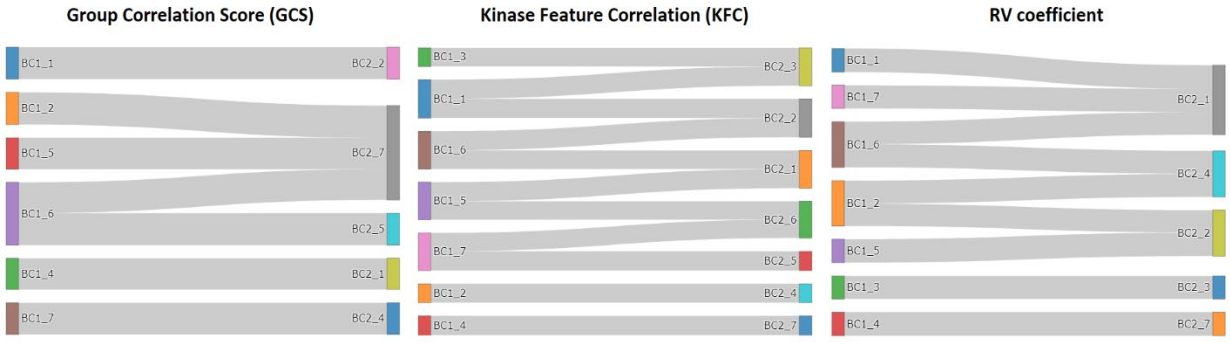


Figure 4. Cluster Overlap Significance between the clusters from BC1 and BC2

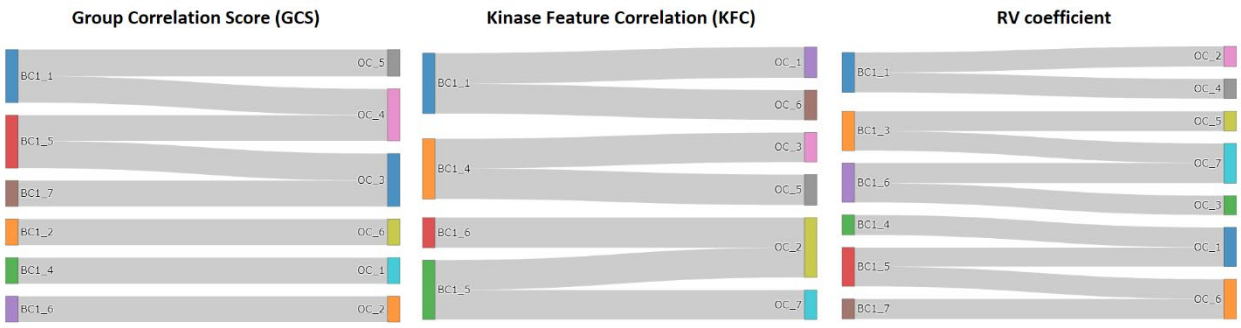


Figure 5. Cluster Overlap Significance between the clusters from BC1 and OC

For discussion, let us consider Figure 3. This cluster sets were computed using the KFC metric we introduced. The data shows that cluster 1 has a significant overlap with the families AGC and CMGC, Additionally, the KFC cluster set for BC2 reported that same combination of families on cluster 1 as well. Looking at Table 1 we observe that while both clusters have AGC, Other, CMGC, CAMK and STE families present, the kinases are almost completely different. For instance, consider the AGC and CMGC families in the BC1_1 and BC2_1 cluster. These clusters have 9 and 8 kinases belonging to AGC respectively. However, PKCH is the only common kinase between them that is a member of the family. Additionally, CMGC is represented by 10 kinases in cluster BC1_1 and 8 in BC2_1 from which 5 are found on both clusters. While this is a larger overlap for CMGC than AGC, the kinases are different, and thus, this 2 clusters do not show up on the cluster overlap analysis for BC1 vs BC2, even though they contain similar distribution of families. In conclusion, these results show some degree of association across the clusters coming from different datasets, suggesting that kinases are grouping together somewhat independent of their families. While we can observe that indeed, the clusters have highly significant overlaps to some of the families, our data shows that clusters that the overlap with a given family do not necessarily share the same kinases.

BC1 – Cluster 1		BC2 – Cluster 2	
KINASE	**FAMILY**	**KINASE**	**FAMILY**
AMPKA1	CAMK	AKT2	AGC
ATM	Atypical	AURA	Other
BRK	TK	BRAF	TKL
CDC2	CMGC	CDC2	CMGC
CDK5	CMGC	CDK4	CMGC
CHK1	CAMK	CHAK1	Atypical
CLK1	CMGC	CK1D	CK1
ERK1	CMGC	COT	STE
ERK5	CMGC	DYRK2	CMGC
GSK3B	CMGC	ERK1	CMGC
JNK2	CMGC	GSK3A	CMGC
LATS1	AGC	HIPK2	CMGC
MAP2K7	STE	IKKB	Other
MAP3K7	STE	MAP2K1	STE
MARK2	CAMK	MAP3K5	STE
MARK3	CAMK	MAP3K7	STE
NEK2	Other	MSK1	AGC
NEK6	Other	NDR1	AGC
P38A	CMGC	P70S6K	AGC
P70S6KB	AGC	PAK1	STE
PKACA	AGC	PAK4	STE
PKCE	AGC	PDK1	AGC
PKCI	AGC	PKCB	AGC
PLK1	Other	PKCI	AGC
RAF1	TKL	PKCZ	AGC
ROCK1	AGC	PKD2	CAMK
RSK2	AGC	PLK3	Other
SGK1	AGC	ROCK2	AGC
		RSK2	AGC

Table 1. Kinases in Cluster 1 of BC1 and Cluster 2 of BC2 (KFC)

Kinase Substrate Enrichment Analysis

Following a similar structure to the previous sections, we analyze the sets of clusters generated by the different metrics for each dataset. As shown in Figure 6, the cluster set for GCS in the BC1 dataset reports some similar average correlations as the individual families that appear in the cluster, while a couple of them had higher values than all the families represented inside. For instance, let us consider cluster 3, which in the previous analysis showed significant overlap with the CMGC (6 members) and Other (5 members) families. The average correlation in those 2 families from the KSEA analysis are .1686 and .1112 respectively, compared to the .3827 value reported by our cluster. The KFC cluster set reports similar results although most of the cluster average correlations are closer to the family values (Supplementary Figure 4).

Looking at cluster 5 we find it to have a mix of different families while producing an average correlation value of .2175, which is slightly above the values reported for the families with the closest being AGC. Finally, the results for the RV coefficient clusters show a similar story with cluster 1, 6 and 7 have significantly higher values than the families inside of them (Supplementary Figure 5).

Continuing with the discussion with BC2, GCS produced cross-correlation means that are in line with the families (Supplementary Figure 6). In general, the results only show cluster 2 being better correlated than the members of the families in it. This cluster features 3 kinases from the CMGC family and 1 from the Other family. KFC reports the pattern where the clusters have similar averages than their family counterparts while having better correlation than the families with which they had a significant overlap in most cases (Supplementary Figure 7). Looking at cluster 4, it reports higher average than CAMK which is the family with whom it had a significant overlap, while also having members of 3 more families. This mix of kinase-families correlates better than the kinases in the family alone. In contrast, cluster 2 did not overlap significantly with any family while having a of kinases from 6 families in total and produced a higher mean correlation than any of those families. Continuing with RV coefficient cluster set, it shows a similar patter with some significant results (Supplementary Figure 8). In general, these results suggest that the phospho-based clusters might represent more functional relevance among kinases as compare to the kinase families.

Finalizing this part of the analysis, let us consider the cluster sets for ovarian cancer. GCS report is in line with the previous results, as the average correlation of the clusters is around the values of the families (Supplementary Figure 9). However, we find some special cases where the clusters either have a noticeable higher average correlation and one with significantly lower. For

instance, looking at cluster 3, which overlapped with families CAMK and STE, reported a lower value than the families inside, containing 3 kinases from CAMK and 1 from STE and AGC respectively. Since these families are better correlated alone, it suggests that the kinases in the cluster are going in different directions in terms of activity. In comparison, cluster 5 has a significantly higher mean correlation than the families, which hints that those kinases from different families combined are following a closer pattern of activation than the kinases from the families alone. For its part, KFC presents a similar picture, where 3 of the clusters captured a combination of kinases from different families which correlate better than the families, highlighting clusters 3, 4, and 6 (Supplementary Figure 10). RV coefficient also present the pattern, these clusters also captured combinations of kinases that correlate better on the perturbation data (Supplementary Figure 11).

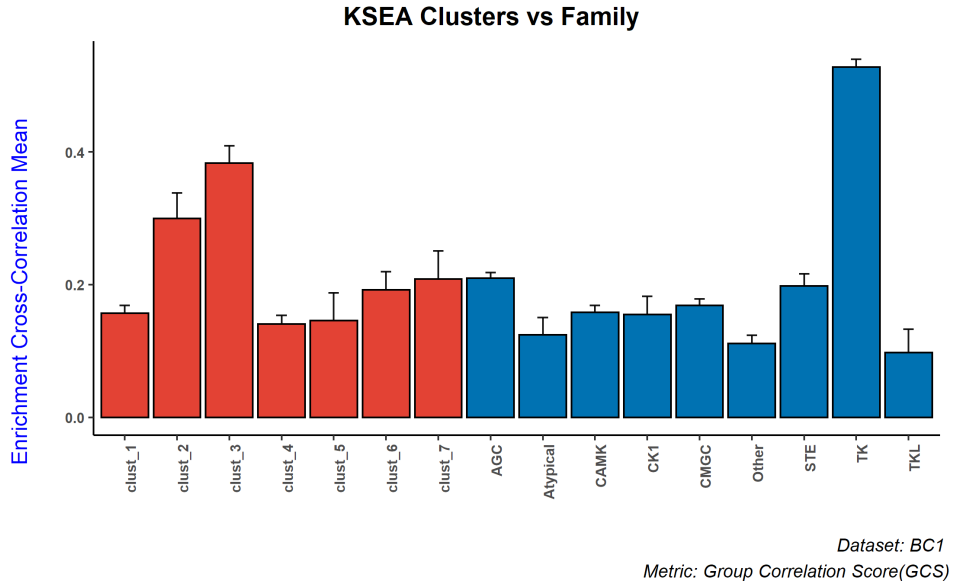


Figure 6. Plot of the cross-correlation mean of the kinase enrichment values

To conclude, while all of our cluster sets did not outperformed the families on every single case, there is enough occurrences to suggest that kinases with similar catalytic domain do not follow the same phosphorylation pattern, even though they may have the same functions in the cell.

Interaction Matrices

In this section, we investigate how kinases from different families are grouped together in the phosphor-based kinase clusters. This analysis would show the collaborative interactions among families.

To open the discussion, we look at the interactions produced by GCS across BC1, BC2 and OC (see Figure 7). The result shows that GCS clusters detected the most common interacting pair of families to be AGC with CAMK, CMGC with Other, and AGC with CMGC, with 5, 6 and 7 interactions across all the datasets respectively. KFC reported the same pairs also to be the most interacting families with 11 and 13 respectively, with the addition of other family combinations such as Other with TK with 9 interactions, and AGC with Other with 8 interactions (Supplementary Figure 12). RV coefficient results are also consistent reporting AGC with CAMK, and CMGC with AGC as the most interacting pairs, while also showing significant interactions of AGC with Other, CMGC with Other (Supplementary Figure 11).

In perspective, we report we found many interactions across the datasets to be rather spread around. However, across all these results we find AGC, CMGC and CAMK to be involved together the most.

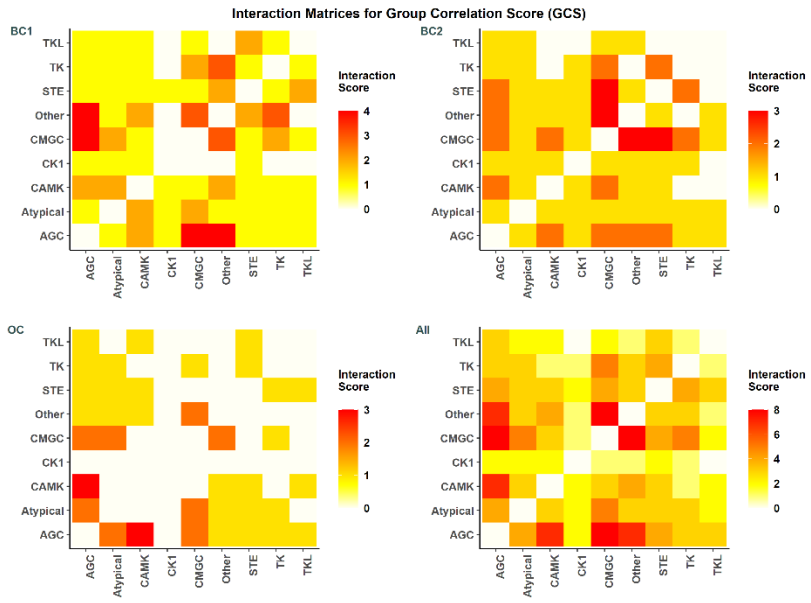


Figure 7. Interaction matrices of the cluster set generated by GCS across the three datasets (BC1, BC2, OC)

Discussion

Kinases are grouped together by the similarity of their catalytic domain which has proved to be a good indicator in the classification of eukaryotic protein kinases [21]. However, many protein kinases have found their catalytic domain to be tethered to one or more non-kinases domains that are responsible for different actions such as regulation, substrate specificity among others. These similarities and differences in the domain architectures indicate some critical features required for functional specialization [22]. Previous research has found the STE group to activate the MAPK family which is part of the set of families enveloped by CMGC [23] which was reported in our interaction matrices. In another study they found CMGC and CAMK, as well as CMGC and AGC families, to be involved in the majority of regulatory circuits of a kinase-kinase interaction network [24]. This is consistent with some of the findings of our interaction matrices.

CHAPTER V

SUMMARY AND CONCLUSION

In this work, we analyze 3 phosphoproteomics datasets, 2 for breast cancer and 1 for ovarian cancer. We proposed 3 different metrics to assess the relationship of kinases based on the correlation of their substrates. Then, we used hierarchical clustering to generate groups of kinases and analyze their relationship to the families found within them by assessing the significant overlap of kinases using the hypergeometric cumulative distribution function. Our results show several combinations of families within clusters suggesting their interaction in the phosphorylation process. We also show the identified clusters are reproducible using independent datasets. We validate the kinase activities of phospho-based clusters by performing kinase enrichment analysis on the perturbation data of phosphosites in 80 different conditions, and we compare the correlation of our clusters in that dataset and the correlation of the families. Our results show combinations of kinase families that had higher correlation than the families of the kinases involved in those clusters.

In general, we have provided different methods to analyze the relationship of kinases and their families and interactions from a Computer Science/Mathematics perspective. These clusters might be useful in identifying alternative drug targets and understanding underlying mechanism and interactions among kinases. Further improvements could be found by extending this analysis include data about their catalytic domain sequences and tethered domains, including data

regarding the kinase families know functions as well as exploring the kinase family relationship in more diseases.

REFERENCES

1. NtwaliP.V, Migisha, et al. “Mass Spectrometry-Based Proteomics of Single Cells and Organoids: The New Generation of Cancer Research.” TrAC Trends in Analytical Chemistry, Elsevier BV, Aug. 2020, p. 116005. Crossref, doi:10.1016/j.trac.2020.116005.
2. Nakagami, Hirofumi, et al. “Large-Scale Comparative Phosphoproteomics Identifies Conserved Phosphorylation Sites in Plants.” Plant Physiology, vol. 153, no. 3, American Society of Plant Biologists (ASPB), May 2010, pp. 1161–1174. Crossref, doi:10.1104/pp.110.157347.
3. Hanks, Steven K., and Tony Hunter. “The Eukaryotic Protein Kinase Superfamily: Kinase (Catalytic) Domain Structure and Classification 1.” The FASEB Journal, vol. 9, no. 8, Wiley, May 1995, pp. 576–596. Crossref, doi:10.1096/fasebj.9.8.7768349.
4. Manning, G. “The Protein Kinase Complement of the Human Genome.” Science, vol. 298, no. 5600, American Association for the Advancement of Science (AAAS), Dec. 2002, pp. 1912–1934. Crossref, doi:10.1126/science.1075762.
5. Ardito, Fatima, et al. “The Crucial Role of Protein Phosphorylation in Cell Signaling and Its Use as Targeted Therapy (Review).” International Journal of Molecular Medicine, vol. 40, no. 2, Spandidos Publications, June 2017, pp. 271–280. Crossref, doi:10.3892/ijmm.2017.3036.
6. Ayati, Marzieh, et al. “CoPhosK: A Method for Comprehensive Kinase Substrate Annotation Using Co-Phosphorylation Analysis.” PLOS Computational Biology, edited

7. byPredrag Radivojac , vol. 15, no. 2, Public Library of Science (PLoS), Feb. 2019, p. e1006678. Crossref, doi:10.1371/journal.pcbi.1006678.
8. Singh, Vishakha, et al. “Phosphorylation: Implications in Cancer.” The Protein Journal, vol. 36, no. 1, Springer Science and Business Media LLC, Jan. 2017, pp. 1–6. Crossref, doi:10.1007/s10930-017-9696-z.
9. Zhang, Qi, et al. “Integrated Proteomics and Network Analysis Identifies Protein Hubs and Network Alterations in Alzheimer’s Disease.” Acta Neuropathologica Communications, vol. 6, no. 1, Springer Science and Business Media LLC, Mar. 2018. Crossref, doi:10.1186/s40478-018-0524-2.
10. Tadesse, Solomon, et al. “Targeting CDK2 in Cancer: Challenges and Opportunities for Therapy.” Drug Discovery Today, vol. 25, no. 2, Elsevier BV, Feb. 2020, pp. 406–413. Crossref, doi:10.1016/j.drudis.2019.12.001.
11. Lee JJ, Loh K, Yap YS. PI3K/Akt/mTOR inhibitors in breast cancer. Cancer Biol Med. 2015;12(4):342-354.
12. Mirk kinase inhibition targets ovarian cancer ascites. (2014). Genes & Cancer, 201. https://doi.org/10.18632/genesandcancer.19
13. Huang, Kuan-lin, et al. “Proteogenomic Integration Reveals Therapeutic Targets in Breast Cancer Xenografts.” Nature Communications, vol. 8, no. 1, Springer Science and Business Media LLC, Mar. 2017. Crossref, doi:10.1038/ncomms14864.
14. Hernandez-Armenta, Claudia, et al. “Benchmarking Substrate-Based Kinase Activity Inference Using Phosphoproteomic Data.” Bioinformatics, edited byJonathen Wren , vol. 33, no. 12, Oxford University Press (OUP), Feb. 2017, pp. 1845–1851. Crossref, doi:10.1093/bioinformatics/btx082.

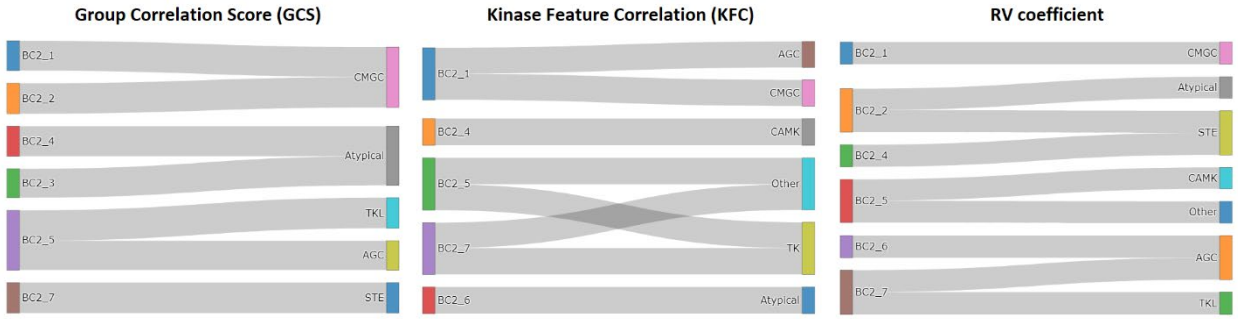
15. Mertins, Philipp, et al. “Proteogenomics Connects Somatic Mutations to Signalling in Breast Cancer.” Nature, vol. 534, no. 7605, Springer Science and Business Media LLC, May 2016, pp. 55–62. Crossref, doi:10.1038/nature18003.
16. Zhang, Hui, et al. “Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer.” Cell, vol. 166, no. 3, Elsevier BV, July 2016, pp. 755–765. Crossref, doi:10.1016/j.cell.2016.05.069.
17. Hornbeck, Peter V., et al. “PhosphoSitePlus, 2014: Mutations, PTMs and Recalibrations.” Nucleic Acids Research, vol. 43, no. D1, Oxford University Press (OUP), Dec. 2014, pp. D512–D520. Crossref, doi:10.1093/nar/gku1267.
18. El Ghaziri, Angélina, and El Mostafa Qannari. “Measures of Association between Two Datasets; Application to Sensory Data.” Food Quality and Preference, vol. 40, Elsevier BV, Mar. 2015, pp. 116–124. Crossref, doi:10.1016/j.foodqual.2014.09.010.
19. Wiredja, Danica D., et al. “The KSEA App: A Web-Based Tool for Kinase Activity Inference from Quantitative Phosphoproteomics.” Bioinformatics, edited byAlfonso Valencia, vol. 33, no. 21, Oxford University Press (OUP), June 2017, pp. 3489–3491. Crossref, doi:10.1093/bioinformatics/btx415.
20. Casado, P., et al. “Kinase-Substrate Enrichment Analysis Provides Insights into the Heterogeneity of Signaling Pathway Activation in Leukemia Cells.” Science Signaling, vol. 6, no. 268, American Association for the Advancement of Science (AAAS), Mar. 2013, pp. rs6–rs6. Crossref, doi:10.1126/scisignal.2003573.
21. Martin, Juliette, et al. “Classification of Protein Kinases on the Basis of Both Kinase and Non-Kinase Regions.” PLoS ONE, edited byJason E. Stajich , vol. 5, no. 9, Public

Library of Science (PLoS), Sept. 2010, p. e12460. Crossref, doi:10.1371/journal.pone.0012460.

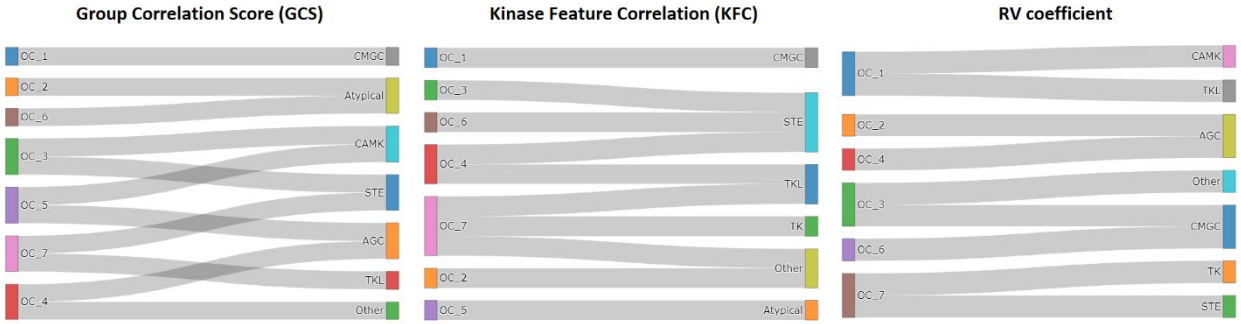
22. Deshmukh, Krupa, et al. “Evolution of Domain Combinations in Protein Kinases and Its Implications for Functional Diversity.” Progress in Biophysics and Molecular Biology, vol. 102, no. 1, Elsevier BV, Jan. 2010, pp. 1–15. Crossref, doi:10.1016/j.pbiomolbio.2009.12.009.
23. Müller-Taubenberger, Annette, et al. “The STE Group Kinase SepA Controls Cleavage Furrow Formation in Dictyostelium.” Cell Motility and the Cytoskeleton, vol. 66, no. 11, Wiley, Nov. 2009, pp. 929–939. Crossref, doi:10.1002/cm.20386.
24. Buljan, Marija, et al. “Kinase Interaction Network Expands Functional and Disease Roles of Human Kinases.” Molecular Cell, vol. 79, no. 3, Elsevier BV, Aug. 2020, pp. 504-520.e9. Crossref, doi:10.1016/j.molcel.2020.07.001.
25. Arshad, Osama A., et al. “An Integrative Analysis of Tumor Proteomic and Phosphoproteomic Profiles to Examine the Relationships Between Kinase Activity and Phosphorylation.” Molecular & Cellular Proteomics, vol. 18, no. 8 suppl 1, American Society for Biochemistry & Molecular Biology (ASBMB), June 2019, pp. S26–S36. Crossref, doi:10.1074/mcp.ra119.001540.

APPENDIX

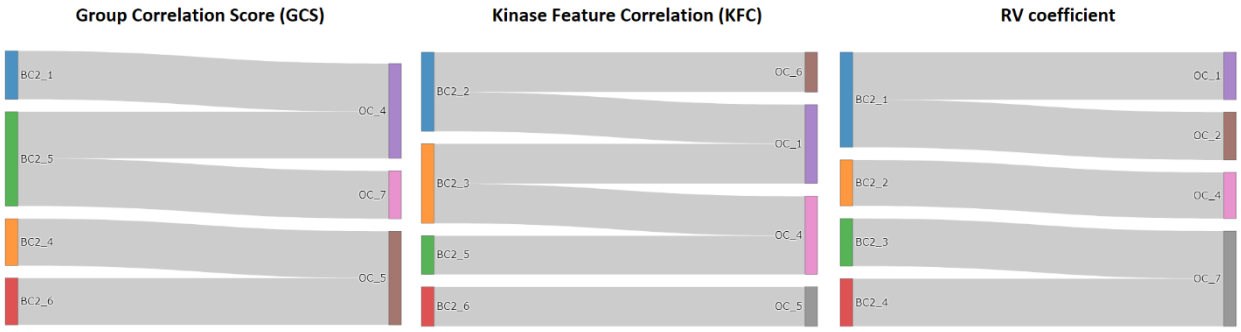
SUPLEMENTARY FIGURES



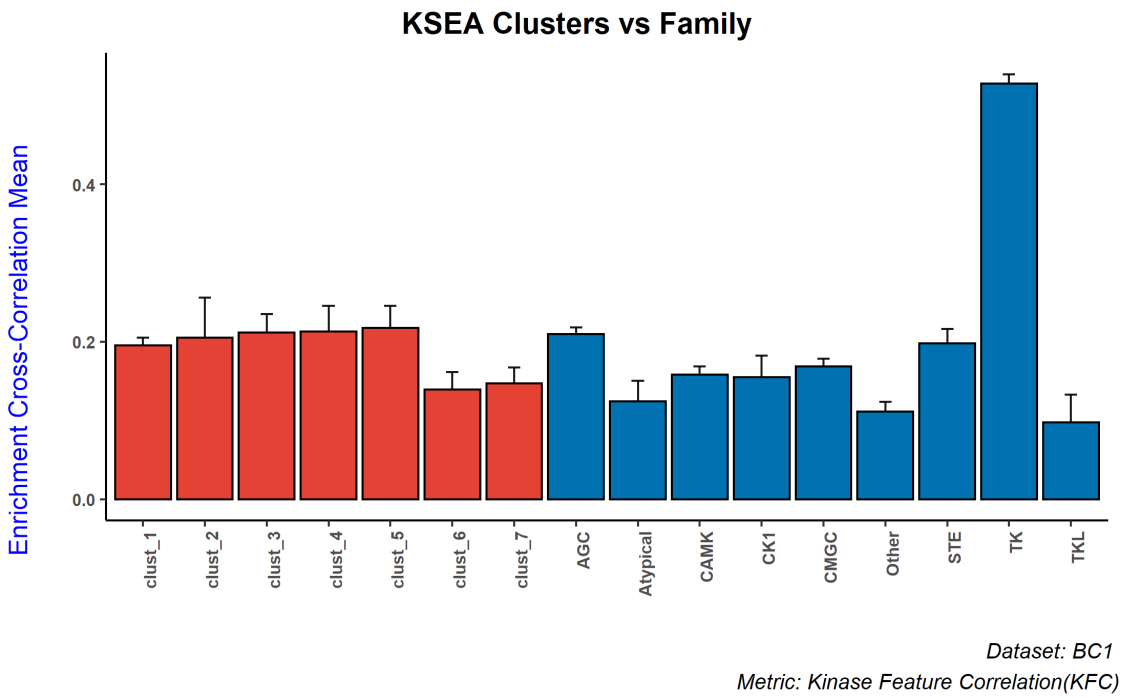
Supplementary Figure 1. Cluster Overlap Significance between the clusters from BC2 and the Family Groups



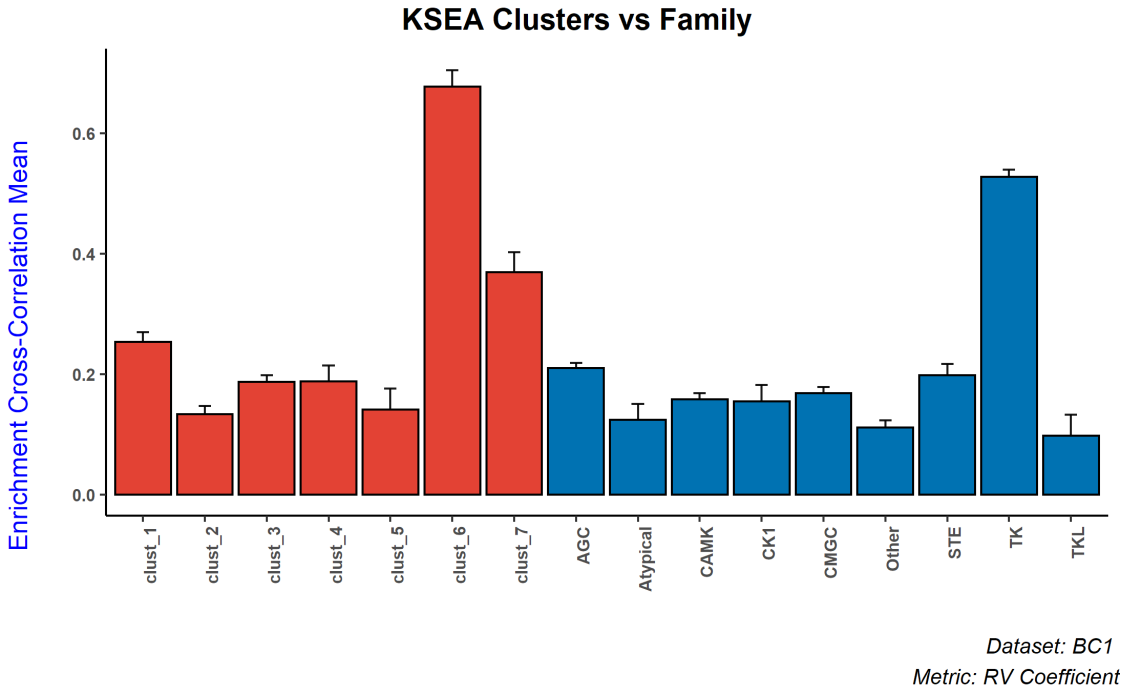
Supplementary Figure 2. Cluster Overlap Significance between the clusters from OC and the Family groups



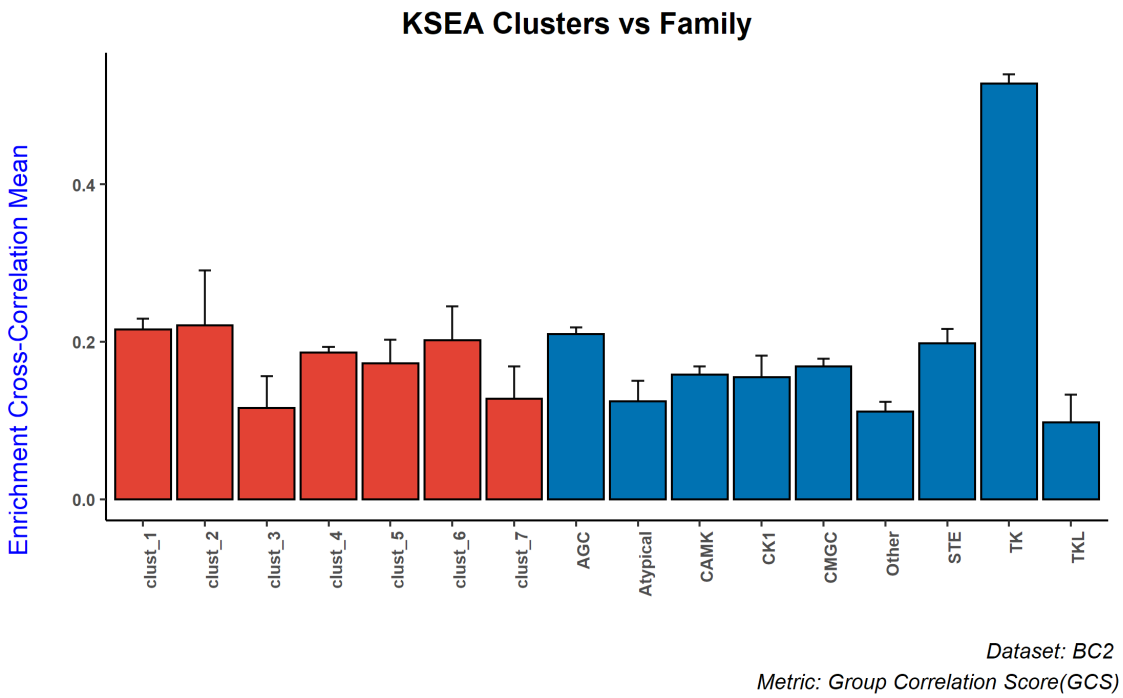
Supplementary Figure 3. Cluster Overlap Significance between the clusters from BC2 and OC



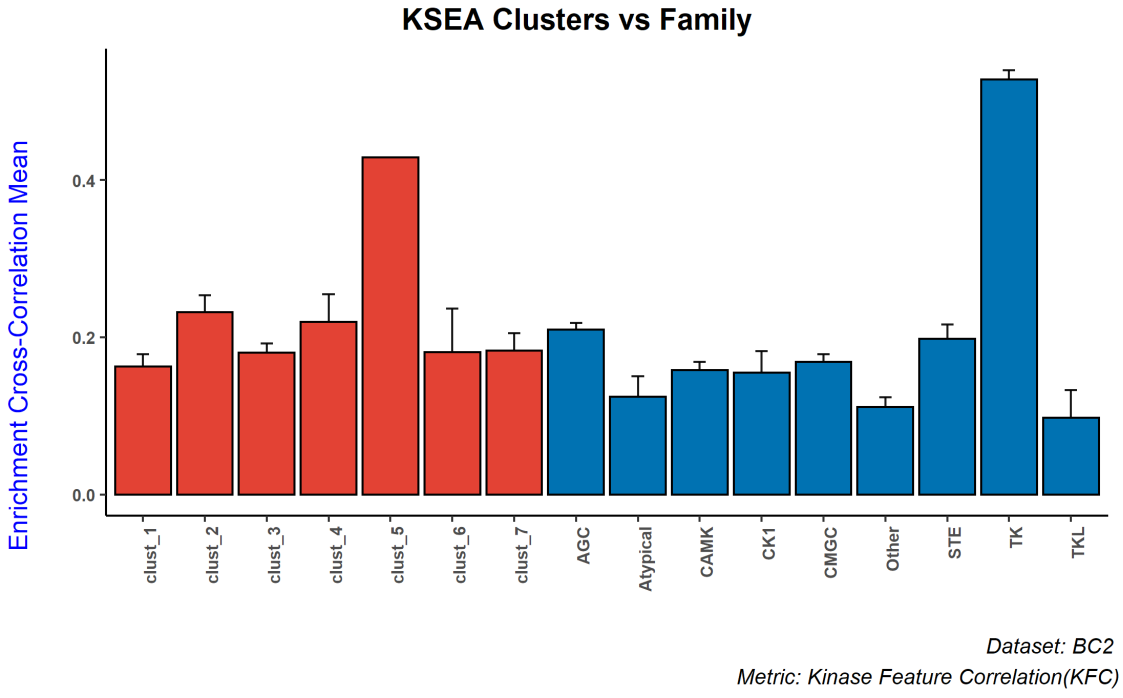
Supplementary Figure 4. Plot of the cross-correlation mean of the kinase enrichment values for BC1 (KFC)



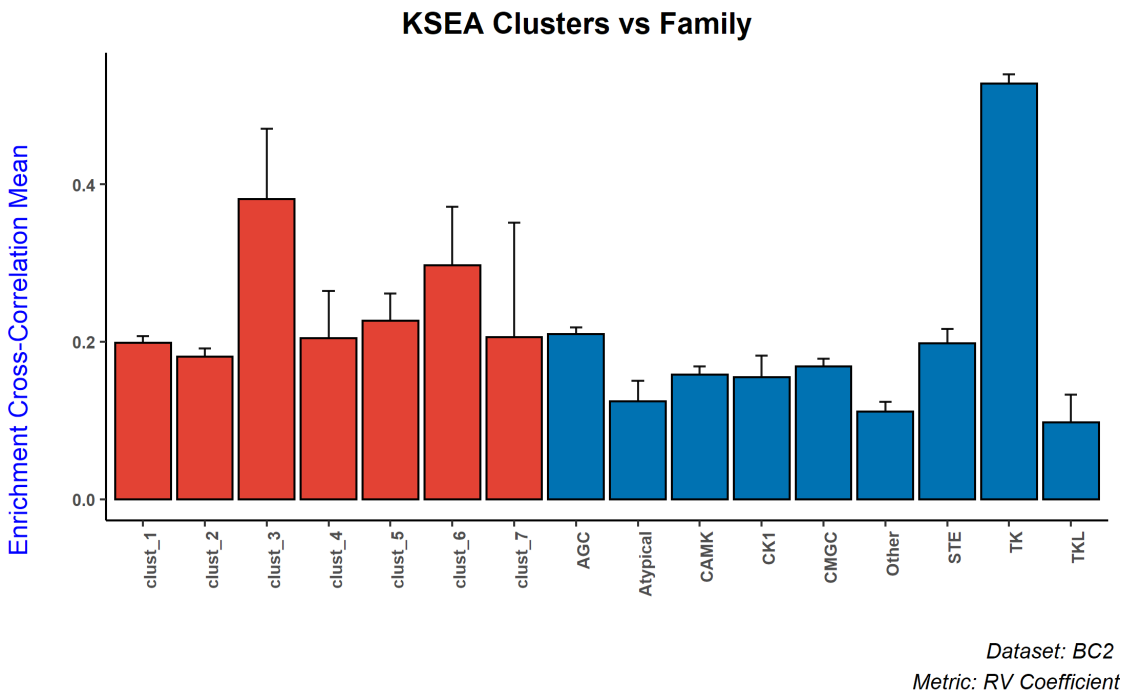
Supplementary Figure 5. Plot of the cross-correlation mean of the kinase enrichment values for BC1 (RV Coefficient)



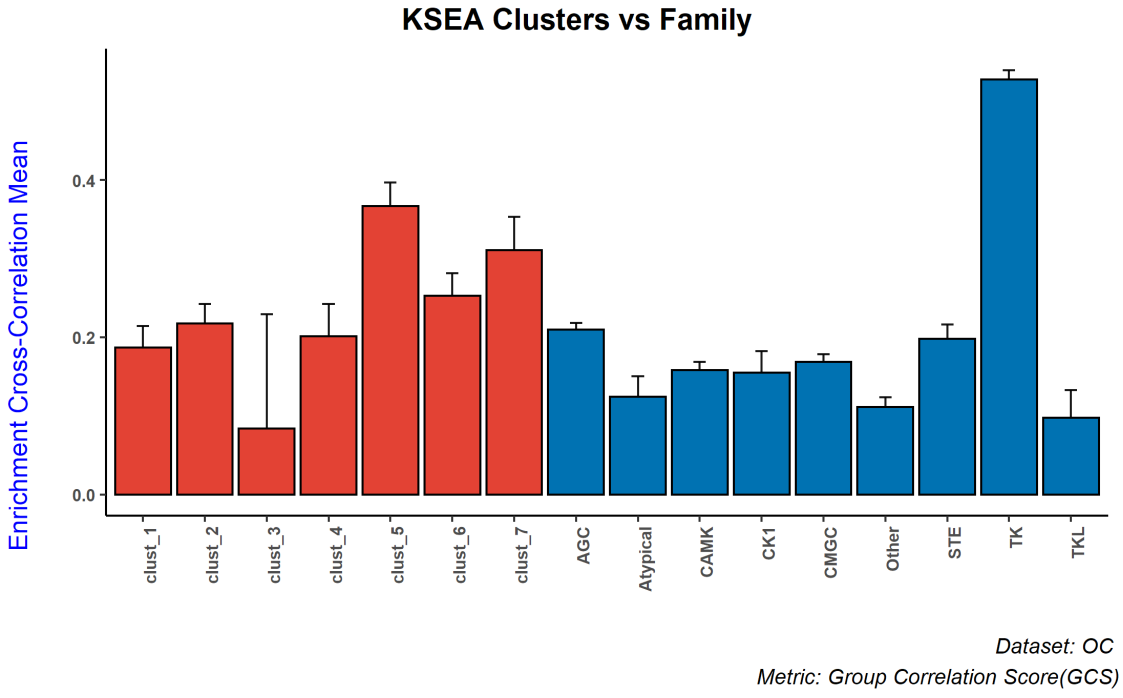
Supplementary Figure 6. Plot of the cross-correlation mean of the kinase enrichment values for BC2 (GCS)



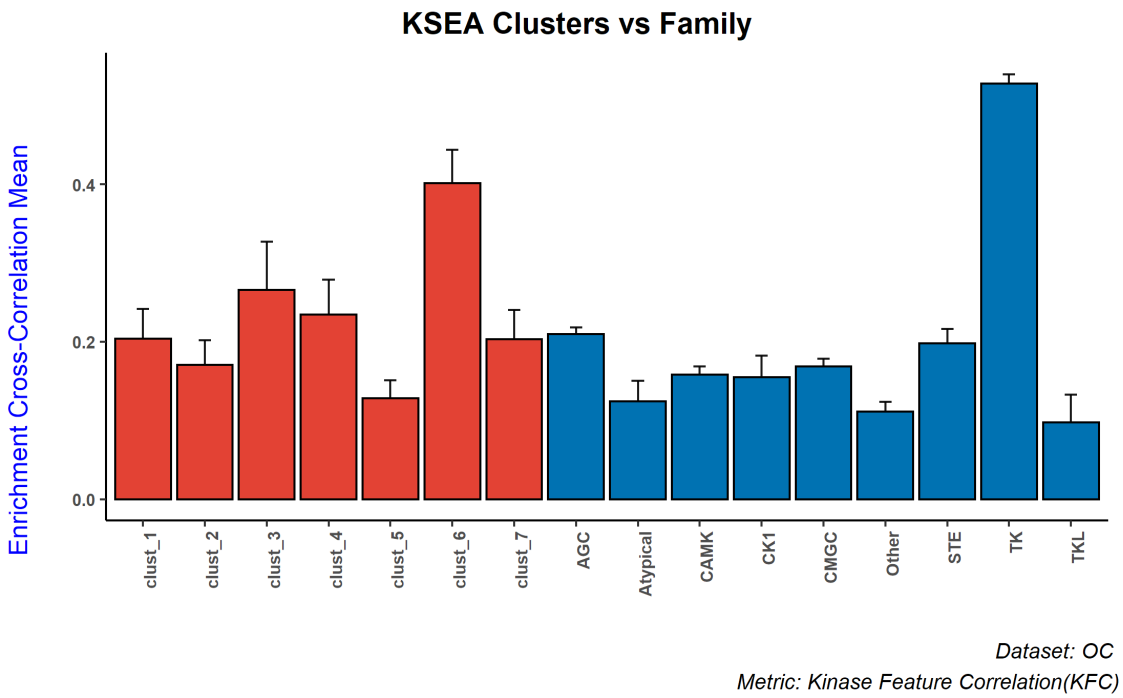
Supplementary Figure 7. Plot of the cross-correlation mean of the kinase enrichment values for BC2 (KFC)



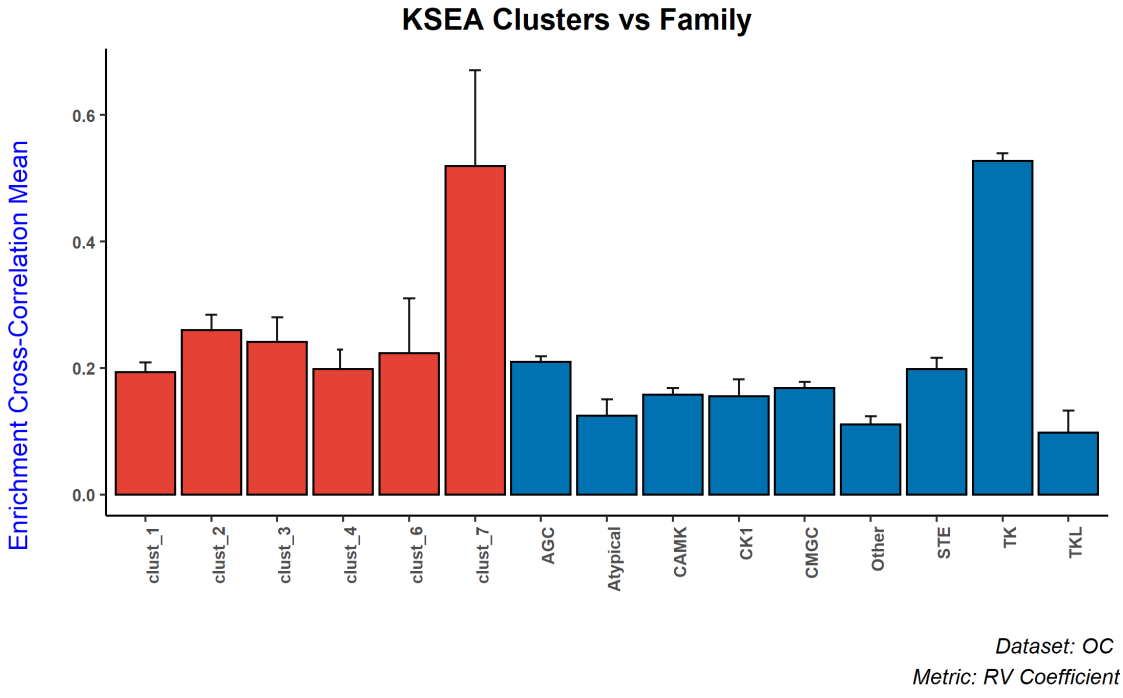
Supplementary Figure 8. Plot of the cross-correlation mean of the kinase enrichment values for BC2(RV Coefficient)



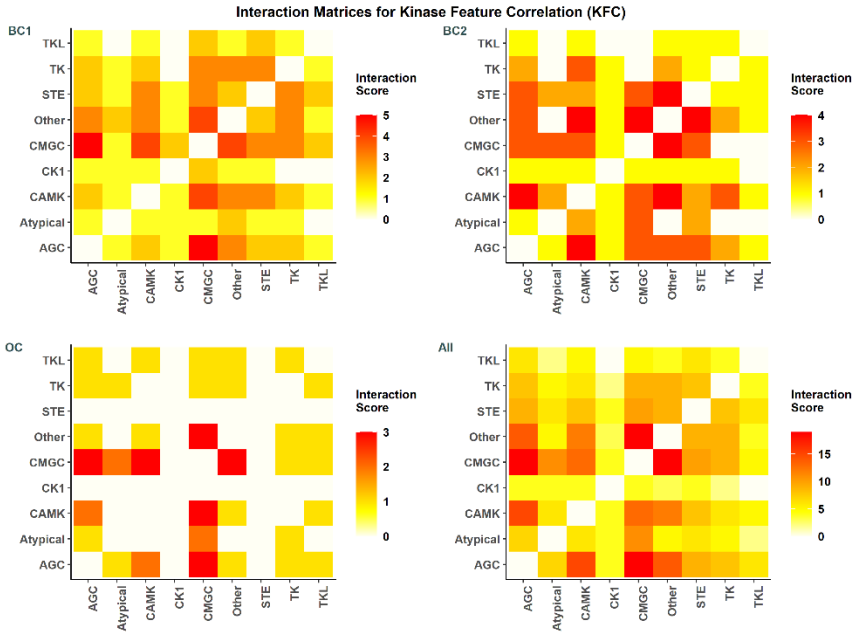
Supplementary Figure 9. Plot of the cross-correlation mean of the kinase enrichment values for OC (GCS)



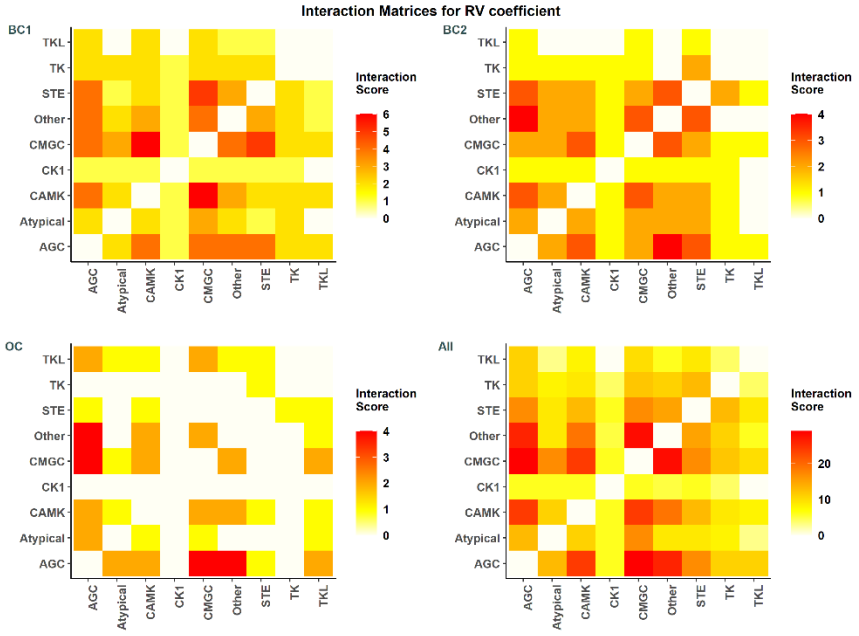
Supplementary Figure 10. Plot of the cross-correlation mean of the kinase enrichment values for OC (KFC)



Supplementary Figure 11. Plot of the cross-correlation mean of the kinase enrichment values for OC (RV Coefficient)



Supplementary Figure 12. Interaction matrices of the cluster set generated by KFC across the three datasets (BC1, BC2, OC)



Supplementary Figure 13. Interaction matrices of the cluster set generated by RV Coefficient across the three datasets (BC1, BC2, OC)

BIOGRAPHICAL SKETCH

David Abdiel Parra Peña, the youngest of two sons, was born in Reynosa, Tamaulipas, Mexico. He attended the Autonomous University of Nuevo Leon (UANL) pursing a degree in Mechatronics before transferring to the University of Texas Pan-American (now UTRGV) where he completed a degree of Science in Computer Engineering in December of 2015. He then earned a Masters of Science in Computer Science from The University of Texas Rio Grande Valley in August 2020. Contact email: davparra91@gmail.com.